강우원, 김다윤, 김소연, 정용헌

NLP 감정분석 기반
마케팅 시장 분석

키워드에 따른 커뮤니티 별 온라인 여론 감정 분석 시스템

빅데이터미네이터

키워드에 따른 커뮤니티 별
온라인 여론 감정 분석 시스템 ☹😐☺

사용자가 입력하는 키워드 기반으로 커뮤니티의 본문 댓글에 대한 크롤링을 진행하고,

이에 따른 다양한 의견들을 감성 분석하여 결과를 도출하고자 합니다.

이는 온라인 여론에 대한 감정 분석을 진행하는 것으로 마케팅 측면에서도 도움이 되는 자료가 될 것입니다.

1. Input

1. 사용자 input
: 키워드

2. 크롤링 범위
: 제목, 본문, 날짜, 댓글

2. 커뮤니티 선정

- 네이트판
- 뽐뿌
- 클리앙
- 루리웹
(+ 네이버 블로그 및 sns 고려)

: 커뮤니티의 순위, 사용자 수,
카테고리의 다양성, 데이터의 양을
중심으로 선정

3. 전처리

- 영어, 특수문자 등 제거
- 초성 (ㅋㅋ, ㅎㅎ, 초성 욕설 등) 제거
- 형태소 단위로 토큰화 → 문장 단위로 순서는 유지해야함
(앞뒤 단어와의 연관성에 따른 감성 분석 가능)
- csv 파일로 저장 (프로토타입)

4. 프로토타입 구조

크롤링부 (+ 전처리)

↓
csv 파일 (추후 DB 구축)

↓ 감성어 사전
(범용 사전 + 감성어 추가)

감성 분석기 => 결과 도출

새롭게 정한 전처리 기준에 맞게 각 커뮤니티에 대한 크롤링부를 구현하고자 합니다.

따라서 각 커뮤니티 별로 크롤러를 구현하는 작업을 시작하였습니다.

커뮤니티 별 크롤링
01. 네이트판

```
In [21]: import os
         import sys
         import pandas as pd
         import requests
         import re
         from datetime import datetime
         from bs4 import BeautifulSoup
         from selenium.webdriver import Chrome
         from selenium.webdriver.common.keys import Keys
         import time
         import json
         import csv

In [22]: browser = Chrome()
         browser.maximize_window()
         base_url = 'https://pann.nate.com'
         browser.get(base_url)
         browser.find_elements_by_xpath('//*[@id="input_search"]')[0].click()

In [23]: #원하는 검색어 입력
         query_txt = input('크롤링할 키워드를 입력하세요: ')
         크롤링할 키워드를 입력하세요: 코로나

In [24]: browser.find_elements_by_xpath('//*[@id="input_search"]')[0].send_keys(query_txt)

In [25]: #검색
         browser.find_elements_by_xpath('//*[@id="search"]/fieldset/button')[0].click()

In [26]: #특정 게시판 검색 결과 더 보기 선택
         browser.find_elements_by_xpath('//*[@id="container"]/div[2]/div[1]/div[3]/p/a')[0].click()
         #추후 최신순 선택 추가

In [35]: nate_page = browser.current_url
         page = nate_page + '&page='
         #page 주소 뒤로 keyword 입력 변수 가져오는 형식으로 바꾸기
         titles = []
         links = []
         #page 1~43까지 추출할때 (10page까지 크롤링시 시간도 다소 걸리며, 본문의 'p'와 'espresso_tditor_view'등의 class에 대한 부분 추가필요)
         for i in range(1, 6):
             res = requests.get(page + str(i))
             res.raise_for_status()
             res.encoding = None
             html = res.text
```

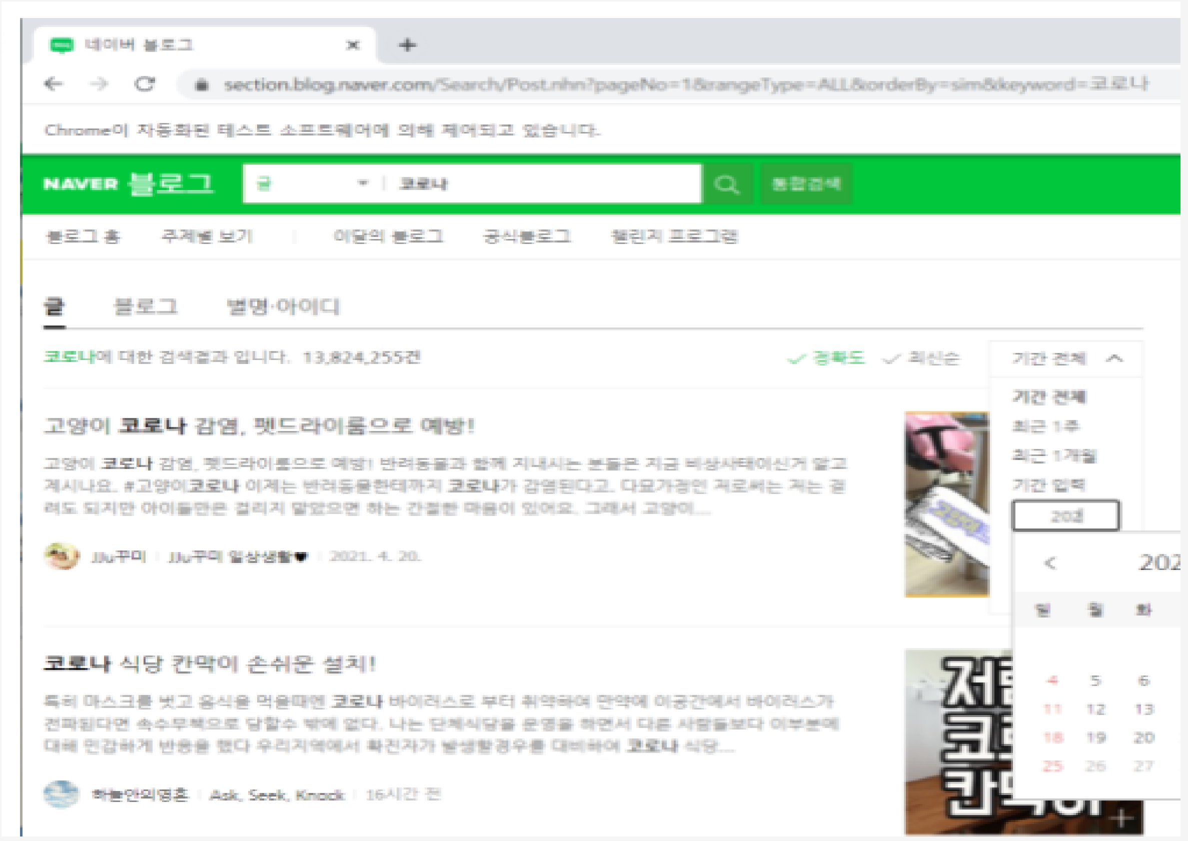
- 키워드를 입력하여 그에 해당하는 '톡톡' 게시판의 글들을 수집합니다. 최신순으로 게시글을 찾고, 게시물 각각에 대한 url을 우선 수집하며, 이 후 개별 url에 대한 본문 및 댓글 크롤링을 진행합니다.
- 긴 본문을 형태소 단위로 토큰화 진행해야 합니다. 또 해당 토큰들이 문장 내의 순서를 유지하도록 하여 추후 앞 뒤 관계에 따른 감성 분석을 진행 하고자 합니다.

커뮤니티 크롤링
02. 클리앙

- 미리 txt 파일에 저장한 키워드에 대한 크롤링이 가능한 상태입니다.
- 프로그램 내에서 키워드 입력에 대한 크롤링으로 수정하고 앞서 회의를 통해 결정한 전처리 양식에 맞추어 크롤링한 데이터에 대한 정리가 필요합니다.

커뮤니티 외 크롤링 Test
03. 네이버 블로그 및 트위터

추후 커뮤니티 외 데이터가 많은 네이버 블로그 혹은 트위터(SNS)에서 감성분석에 적합한 데이터를 많이 얻을 수 있는지 test를 해 보았습니다.



▪ 데이터의 양이 매우 많고 또 최신 데이터가 많다는 장점이 있지만, 광고의 양도 많다는 단점이 있습니다.
=> 데이터 처리를 잘 한다면 활용가능한 자료가 될 것입니다.

			username	tweet	retweets_count
88	2021-04-13	08:26:29	mbcnews	지난해에는 코로나19로 행사가 취소됐지만 올해는 예정대로 진행돼 수백 명의 참가자가...	10
89	2021-04-13	08:26:04	bissori0613	@CtWwBTS1 바다야 거기 사는 모두에게 우리가 미안해 코로나도 중요하지만 ...	0
90	2021-04-13	08:25:41	didchdkfl	코로나19에는 비대면 핸선라이프~ "말씀광장" 과 함께 핵 사이다 말씀 맛집 gog...	0
91	2021-04-13	08:25:19	jason326naverc2	코로나ㅏㅏ...	0
92	2021-04-13	08:24:54	kt114	논산여자중학교 제24기 운영위원회가 교목인 소나무를 기념식수 하며 출발했네요. 최병...	0

▪ 자주 검색되는 단어의 경우 사람들이 노출 수를 올리기 위해 일부러 해시태그에 포함하기 때문에 관련 없는 트윗까지 크롤링 된다는 단점이 있습니다.
=> 주어지는 본문 글이 짧다는 단점이 있어 비교적 활용성이 떨어진다고 판단했습니다.

+

차주에는

1. 선정한 커뮤니티 4곳에 대한 크롤링부 (전처리를 포함한) 심화

2. 크롤링 결과에 따라 어떤 감성어를 추가할 것인지 분석

3. 감성어 사전 및 감성 분석 함수에 대한 학습 및 고민

을 진행하고자 합니다.

감사합니다 :)

빅데이터미네이터

손윤식 교수님 주식회사 닥터송 송환구 멘토님